Proc. Natl. Acad. Sci. USA
Vol. 75, No. 8, pp. 3737–3741, August 1978
Biochemistry

Nucleotide sequence of the ampicillin resistance gene of *Escherichia coli* plasmid pBR322

(protein sequence/secretion signal/β-lactamase/DNA chemistry)

J. GREGOR SUTCLIFFE*

The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138

Communicated by Walter Gilbert, June 14, 1978

ABSTRACT I have determined the nucleotide sequence of the ampicillin resistance gene of pBR322, an *Escherichia coli* plasmid that encodes a penicillin β-lactamase. This gene codes for a protein of 286 amino acid residues. The first 23 amino acids presumably form a signal for secretion, because they do not appear in the mature enzyme, whose partial amino acid sequence has been determined independently, [Ambler, R. P. & Scott, G. K. (1978) *Proc. Natl. Acad. Sci. USA* 75, 3732–3736].

Recent advances in direct DNA sequencing methods have greatly increased the scope of the DNA sequences that are accessible. As large amounts of sequence data are accumulated, it is important to know how reliable the data are likely to be, as well as how quickly a given sequence may be determined. The ampicillin resistance (*amp*^{r}) gene provided a sequence whose accuracy could be tested;† determining this sequence was also a desirable goal in its own right.

The *amp*^{r} gene of *Escherichia coli* and other Gram-negative bacteria codes for a β-lactamase (penicillin amido-β-lactamhydrolase, EC 3.5.2.6) of approximately 27,000 daltons that catalyzes the hydrolysis of penicillins to penicilloic acids. Rational design of penicillins that are resistant to this penicillinase requires the elucidation of the structure of the enzyme's active site. To this end a crystallographic study of the protein is in progress (1), and mutations that alter the kinetic properties of the enzyme (presumably by altering the active site) have been generated (2). The amino acid sequence of β-lactamase is essential to solving the structure of the enzyme.

The *amp*^{r} gene is easily accessible because it is carried along with tetracycline resistance on the plasmid vector pBR322. pBR322 is superior to other cloning vectors, and any information about this plasmid is likely to be useful, especially because pBR322 has been approved by the National Institutes of Health as an EK2 vector for cloning in *E. coli.* The plasmid R1 (originally designated R7268) was isolated from the wild in London (1963) from *Salmonella paratyphi B* (3). The *amp*^{r} gene was transposed to pBR312 from R1 via pSF2124 (4–6). Molecular rearrangements of pBR312 produced pBR322 (7). The purified β-lactamase of R1 has been shown to be of type RTEM 1 (8).

Plasmid DNA was prepared by chloramphenicol amplification (9) in hosts RR1 (lacking restriction and modification specificity of K-12) (6) or GM119 (*dam-3*) (10). The *amp*^{r} gene was thought to cover the unique site for the restriction enzyme *Pst* I (6, 7), and I confirmed this by opening the DNA circle with *Pst* I, then treating the opened plasmid sequentially with S1 nuclease and DNA ligase. The single-strand DNase S1 was expected to chew back the tetranucleotide 3′ extension characteristic of *Pst* I cleavage (11), and the ligase was expected to reseal the circle. The DNA so treated was used to transform *E. coli*, and of 26 transformants selected for tetracycline resistance, 20 were found to be ampicillin sensitive. A lesion at the *Pst* I site, therefore, inactivates the penicillinase. Thus, at the onset of this project I had two pieces of information: (*i*) that the *amp*^{r} gene covered the unique *Pst* I site, and (*ii*) that the size of the β-lactamase protein was about 27,000 daltons. This meant I needed a sequence of at least 729 base pairs overlapping the *Pst* I site.

The plasmid has unique restriction sites for the enzymes *Eco*RI and *Sal* I (as well as *Pst* I) and, in addition, two sites for the enzyme *Hin*dII, one of which is the same as the *Sal* I site (7). The *Eco*R1 site is about 750 nucleotide pairs from the *Pst* I site (determined by gel electrophoresis); the *Sal* I site is remote from the *Pst* I site; and the other *Hin*dII site lies about midway between the *Eco*RI and *Pst* I site. I compared restriction digests generated by one of the frequently cutting enzymes (*Hae* III, *Hpa* II, *Alu* I, *Hin*fI, *Taq* I, *Tha* I) with double digests generated by that enzyme and a rarely cutting enzyme (*Eco*RI, *Sal* I, *Pst* I, *Hin*dII), and thereby determined which restriction fragments lie between the *Eco*RI and *Pst* I sites. The largest *Taq* I fragments of pBR322 contained the *Pst* I site, and this was examined to determine which sites for other restriction enzymes it contained. When the enzyme *Mbo* I was used, the DNA had to be prepared in the host GM119, which lacks the deoxadenosine methylase (10). *Mbo* I cuts unmethylated but not methylated DNA.

These results were sufficient for making a crude restriction map of the *amp*^{r} gene. I selected relevant fragments and sequenced them, using the method of Maxam and Gilbert (12). Fig. 1 is an example of an autoradiogram, which is the data from which a sequence is directly read. I ensured the continuity of all adjacent restriction fragments by sequencing across all junctions on overlapping fragments. All sequences were determined several times and, whenever it was experimentally convenient, both strands were sequenced. Fig. 2 shows the sequences that I determined.

The Sequence. Fig. 3 represents the sequence of the DNA from the unique *Eco*RI site through the end of the structural gene for β-lactamase. At position 750–755 is the sequence C-T-G-C-A-G, the recognition sequence of *Pst* I. Of the six pos-

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

* Current Address: Department of Cellular and Developmental Immunology, Scripps Clinic and Research Foundation, 10666 North Torrey Pines Road, La Jolla, CA 92037.

† This project was first considered on Feb. 8, 1977, and I began sequencing in March. Soon thereafter R. P. Ambler and G. K. Scott sent their partial amino acid sequence data for the penicillin β-lactamase of *E. coli* to Jeremy Knowles at Harvard University, who held the data until the final DNA sequence was presented to him. On Sept. 8, I considered the data to be unambiguous and presented them to Walter Gilbert, who, after interpreting a subset of the autoradiograms, concurred with the sequence. The comparison of the DNA sequence with the partial amino acid sequence occurred at tea on Sept. 25, 1977.

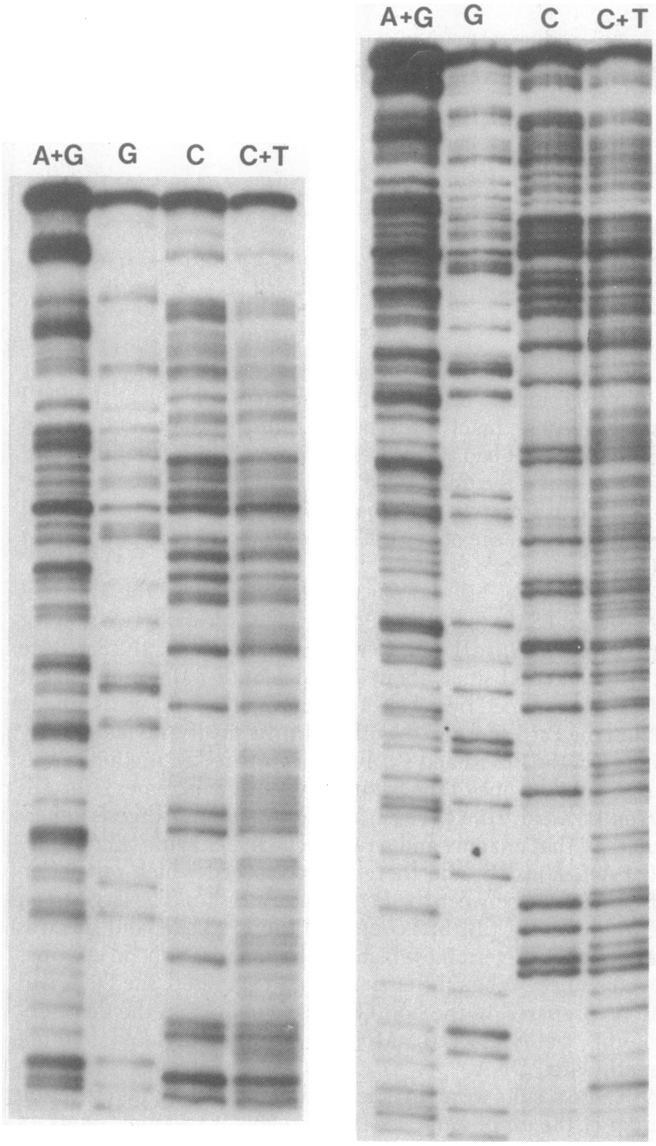

FIG. 1. Maxam–Gilbert sequence ladder from *amp*^{r} gene. This represents an autoradiogram of a polyacrylamide slab gel. The gel (20% acrylamide/0.67% bisacrylamide/7 M urea/50 mM Tris borate (pH 8.3)/1 mM EDTA) was run for 40 hr at 1300 V. The loading of the four lanes on the right side of the photograph was delayed 12 hr. The ^{32}P- labeled 5′ end for this sequence run is the top strand *Hin*fI site at position 996 (see Fig. 3). The reactions used were as described in ref. 12 except that magnesium acetate was omitted from the hydrazine stop solutions. Reactions are, from left, A > G, alternate G only, C only, and C + T. The first 40 bases have been run off the bottom of the gel in the left lanes. The pattern can easily be read further than 150 bases from the 5′ end (the A > G track is better resolved in lighter exposures). The alternate G only reaction was used because it decreases the ambiguity that confuses the G > A reaction at purines distal to the 5′ end. The light ghost purine bands that show up in the pyrimidine tracks are routinely observed, but in general are not a source of confusion. It can be seen that the chemistry is working fine 150 bases from the end, but that this particular gel system no longer resolves well. Variations in electrophoresis conditions can increase the resolution such that autoradiograms can be read 200–300 bases from the 5′-end label (A. Maxam, R. Tizard, G. Sutcliffe, and W. Gilbert, unpublished results).

sible translation reading frames (three in each orientation) that cover the *Pst* I site, only one frame can code for a hypothetical polypeptide of appropriate length without encountering an in-phase nonsense codon. The amino acids encoded in this frame are also indicated in Fig. 3. The first in-phase initiator codon (ATG) corresponds to the first methionine residue. A continuous polypeptide of 286 amino acids can be read, terminating at an ochre codon (TAA). The sequence of the *amp*^{r} gene was determined in the absence of any knowledge of protein sequence.

The amino terminus of purified β-lactamase was determined to be His-Pro-Glu-Thr-Leu by Alan Hall (personal communication). This sequence matches the residues 24–28 of the hypothetical protein decoded from the DNA sequence. The partial protein sequence of Ambler and Scott (13) totally agrees with the DNA sequence, with one exception. The details of that comparison are chronicled in the accompanying paper, along with the details of the protein sequencing, but a few features will be discussed here. The Gln-37 of the pBR322 β-lactamase differs from Lys at that position in the Edinburgh sequence. It is likely that both assignments are correct. In the pBR322 case a single base change (CAG → AAG) would allow Gln → Lys; however, such a change in the DNA sequence would remove a restriction site for the enzyme *Mbo* I (G-A-T-C → G-A-T-A). The restriction site is definitely present, as is demonstrated by overlapping sequence as well as by sequence using the 5′ ends generated by the enzyme. Also, I have shown that *Mbo* I will not cut at G-A-T-A (such a site at position 175 is not cut by *Mbo* I, while the G-A-T-C site at 315 is cut). The Edinburgh sequence comes from the enzyme expressed by the plasmid R6K. The Lys must be correct because it specifies the end of a tryptic fragment. If both assignments are correct, the pBR322 and R6K proteins should differ in pI.[‡]

Glutamic acid/glutamine and aspartic acid/asparagine differences are hard to pick up by protein sequencing techniques, as are tryptophan residues. The DNA sequencing does not have these difficulties. The potential problems in the interpretation of DNA sequencing data are independent of those of protein sequencing.

The result of the comparison of the hypothetical protein predicted by the DNA sequence and the partial protein sequence is that the DNA sequence is probably totally correct. Errors, it they exist, are confined to degeneracies of the genetic code. The *amp*^{r} gene sequence was solved without consulting the protein data so that it could be determined if the DNA sequence analysis could stand alone. A distance of 789 nucleotides with no detectable errors shows that, in fact, distance is no obstacle for sequence determination. Confidence can be placed in carefully sequenced regions, even if no cross-check with protein is available. As a corollary, if some data about a protein, such as the composition of its tryptic fragments, is available, then DNA sequence may be even more easily obtained. Fewer than 7 mo were required both to learn all the techniques of DNA sequencing and to obtain the 1100 base pairs of confirmed sequence presented here.

[‡] This point is the source of some confusion. R1 and R6K penicillinases have each been classified as RTEM 1 by isoelectric focusing. Two classes exist, TEM 1 having a pI of 5.4 and TEM 2 of 5.6. Either the pBR322 penicillinase has an even lower pI than RTEM 1 or else there has been a mix-up in the history of R6K. A. Hall and J. Knowles (personal communication) report that their "R6K," obtained from the Microbiological Research Establishment, Porton, England (which also supplied Ambler and Scott, and Knox's group), carries resistance to tetracycline, but not to streptomycin. These phenotypes differ from those reported by Heffron *et al.* (14). I suggest that the pBR322 enzyme is RTEM 1 and the "R6K" enzyme of Ambler and Scott, Hall and Knowles, and probably Knox *et al.* is actually RTEM 2. The difference between TEM 1 and 2 would be explained by the Gly/Lys difference discussed in the text. Because the two types of enzyme have the same activity spectrum against various penicillins and cephalosporins, the altered residue must not be involved in the enzymatic activity.

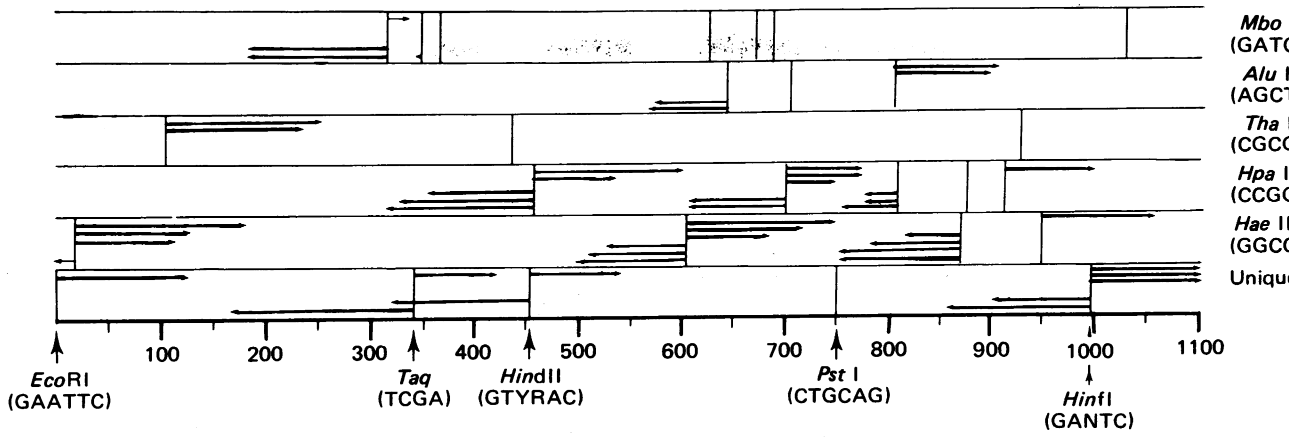

FIG. 2. Restriction map and sequencing strategy for pBR322 *amp*^{r} gene. The numbers along the abscissa represent base pairs and correspond to the DNA sequence of Fig. 3. Each horizontal strip represents the cleavage map for a different restriction enzyme; that enzyme and its cutting sequence are indicated at the end of the strip. The bottom strip (labeled unique) represents the map for enzymes that cut only once in the *amp*^{r} region. The arrows within the strips show the extent of individual sequencing runs. The tail of the arrow is at the 5′ end of the sequenced fragment. Two features of the sequencing allowed complete confidence in the cumulative results. Seventy percent of the gene sequence was determined on both strands. Occasional regions of autoradiograms show peculiar band-to-band spacing, a phenomenon that can be accounted for by intramolecular secondary structure of the DNA during electrophoresis. When both strands have been sequenced, the possibility for errors in film reading is reduced because even if both strands show the peculiarity, it will be located at opposite regions of the sequence such that each portion of the DNA is clearly read on at least one strand. The second feature is that the films were subjected to multiple readings. It is the nature of the data that they consist of long strings of meaningless symbols. Clerical errors can be rampant.

It is a feature of DNA that it has more to say than just what protein it encodes. The first 23 amino acids predicted by the DNA sequence do not appear in the mature, secreted protein. These 23 amino acids are predominantly hydrophobic. If both transcription and translation are colinear with the DNA, then the β-lactamase carries a "leader" of 23 amino acids. The signal hypothesis of Blobel's group (15) ascribes the ability of proteins that must pass through or become part of the cell membrane to the presence of a hydrophobic amino terminus on nascent polypeptides. This signal functions in the processes of transport and secretion and then is cleaved from the primary translation product to form the active molecule. The RTEM β-lactamase fits nicely with this hypothesis. The genetic information for the signal is contiguous with that for the structural enzyme. It is not known if the signal functions by virtue of specific features of its sequence or because of the character that it endows on the rest of the protein. The observation that some secreted proenzymes (e.g., *Bacillus licheniformis* penicillinase) (16) have full enzymatic activity argues that the signal need not inhibit this activity. Presumably the removal of the signal relegates the protein to the external cell compartment. The penicillinase on pBR322 should provide a good model system for using recombinant DNA technology to study the compartmentalization of prokaryotic proteins.

As more gene sequences become known, a picture will emerge as to how the genetic code is utilized by the organism. Table 1 shows the distribution of codons used to assemble the amino acids of β-lactamase. Although the source of this gene in the wild was *Salmonella*, it has functioned well in laboratory *E. coli* for over 10 years. A few asymmetries in codon usage appear for *amp*^{r}. However, when the composite of all available genes for which a sequence is known [*lac i*, *cro*, and the proteins of phages φX174 and Ms2 (17–20)] is surveyed, the only noticeable asymmetry is in arginine codons. CGY is used more frequently than CGR or AGR.

Table 1. Codon distribution for pBR322 β-lactamase

Phe	TTT	9	Ser	TCT	3	Tyr	TAT	1	Cys	TGT	1
Phe	TTC	1	Ser	TCC	3	Tyr	TAC	3	Cys	TGC	2
Leu	TTA	5	Ser	TCA	2		TAA	—		TGA	—
Leu	TTG	4	Ser	TCG	1		TAG	—	Trp	TGG	4
Leu	CTT	9	Pro	CCT	2	His	CAT	4	Arg	CGT	7
Leu	CTC	2	Pro	CCC	3	His	CAC	3	Arg	CGC	7
Leu	CTA	5	Pro	CCA	6	Gln	CAA	5	Arg	CGA	2
Leu	CTG	8	Pro	CCG	3	Gln	CAG	4	Arg	CGG	1
Ile	ATT	5	Thr	ACT	6	Asn	AAT	3	Ser	AGT	5
Ile	ATC	7	Thr	ACC	3	Asn	AAC	5	Ser	AGC	2
Ile	ATA	5	Thr	ACA	4	Lys	AAA	6	Arg	AGA	2
Met	ATG	10	Thr	ACG	7	Lys	AAG	5	Arg	AGG	0
Val	GTT	6	Ala	GCT	8	Asp	GAT	11	Gly	GGT	8
Val	GTC	2	Ala	GCC	8	Asp	GAC	5	Gly	GGC	4
Val	GTA	5	Ala	GCA	9	Glu	GAA	11	Gly	GGA	4
Val	GTG	2	Ala	GCG	4	Glu	GAG	9	Gly	GGG	5

The triplet codons for the amino acids are represented 5′→3′ in DNA. The numbers represent how many times the particular triplet appears in phase in the DNA sequence. AGG is never used for β-lactamase. TAA, TAG, and TGA are nonsense codons whose appearance would indicate polypeptide chain termination. TAA is the single terminator used for β-lactamase.

```
GAATTCTTGAAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTG
                      20                  40                  60                  80                 100
CTTAAGAACTTCTGCTTTCCCGGAGCACTATGCGGATAAAAATATCCAATTACAGTACTATTATTACCAAAGAATCTGCAGTCCACCGTGAAAAGCCCCTTTACAC


    CGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAG
                    120                 140                 160                 180                 200
    GCGCCTTGGGGATAAACAAATAAAAAGATTTATGTAAGTTTATACATAGGCGAGTACTCTGTTATTGGGACTATTTACGAAGTTATTATAACTTTTTCCTTC


                                    10                            20                            30
      MetSerIleGlnHisPheArgValAlaLeuIleProPhePheAlaAlaPheCysLeuProValPheAlaHisProGluThrLeuValLysValLysAsp
    AGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGAT
                    220                 240                 260                 280                 300
    TCATACTCATAAGTTGTAAAGGCACAGCGGGAATAAGGGAAAAAACGCCGTAAAACGGAAGGACAAAAACGAGTGGGTCTTTGCGACCACTTTCATTTTCTA


                  40                            50                            60
    AlaGluAspGlnLeuGlyAlaArgValGlyTyrIleGluLeuAspLeuAsnSerGlyLysIleLeuGluSerPheArgProGluGluArgPheProMetMet
    GCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATG
                 320                 340                 360                 380                 400
    CGACTTCTAGTCAACCCACGTGCTCACCCAATGTAGCTTGACCTAGAGTTGTCGCCATTCTAGGAACTCTCAAAAGCGGGGCTTCTTGCAAAAGGTTACTAC


          70                            80                            90                            100
    SerThrPheLysValLeuLeuCysGlyAlaValLeuSerArgValAspAlaGlyGlnGluGlnLeuGlyArgArgIleHisTyrSerGlnAsnAspLeuVal
    AGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTT
              420                 440                 460                 480                 500
    TCGTGAAAATTTCAAGACGATACACCGCGCCATAATAGGGCACAACTGCGGCCCGTTCTCGTTGAGCCAGCGGCGTATGTGATAAGAGTCTTACTGAACCAA


                            110                           120                           130
    GluTyrSerProValThrGluLysHisLeuThrAspGlyMetThrValArgGluLeuCysSerAlaAlaIleThrMetSerAspAsnThrAlaAlaAsnLeu
    GAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTA
           520                 540                 560                 580                 600
    CTCATGAGTGGTCAGTGTCTTTTCGTAGAATGCCTACCGTACTGTCATTCTCTTAATACGTCACGACGGTATTGGTACTCACTATTGTGACGCCGGTTGAAT


              140                           150                           160
    LeuLeuThrThrIleGlyGlyProLysGluLeuThrAlaPheLeuHisAsnMetGlyAspHisValThrArgLeuAspArgTrpGluProGluLeuAsnGlu
    CTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAA
        620                 640                 660                 680                 700
    GAAGACTGTTGCTAGCCTCCTGGCTTCCTCGATTGGCGAAAAAACGTGTTGTACCCCCTAGTACATTGAGCGGAACTAGCAACCCTTGGCCTCGACTTACTT


    170                           180                           190                           200
    AlaIleProAsnAspGluArgAspThrThrMetProAlaAlaMetAlaThrThrLeuArgLysLeuLeuThrGlyGluLeuLcuThrLeuAlaSerArgGln
    GCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAA
          720                 740                 760                 780                 800
    CGGTATGGTTTGCTGCTCGCACTGTGGTGCTACGGACGTCGTTACCGTTGTTGCAACGCGTTTGATAATTGACCGCTTGATGAATGAGATCGAAGGGCCGTT


                        210                           220                           230
    GlnLeuIleAspTrpMetGluAlaAspLysValAlaGlyProLeuLeuArgSerAlaLeuProAlaGlyTrpPheIleAlaAspLysSerGlyAlaGlyGlu
    CAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAG
     820                 840                 860                 880                 900
    GTTAATTATCTGACCTACCTCCGCCTATTTCAACGTCCTGGTGAAGACGCGAGCCGGGAAGGCCGACCGACCAAATAACGACTATTTAGACCTCGGCCACTC


          240                           250                           260                           270
    ArgGlySerArgGlyIleIleAlaAlaLeuGlyProAspGlyLysProSerArgIleValValIleTyrThrThrGlySerGlnAlaThrMetAspGluArg
    CGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGA
    920                 940                 960                 980                 1000                1020
    GCACCCAGAGCGCCATAGTAACGTCGTGACCCCGGTCTACCATTCGGGAGGGCATAGCATCAATAGATGTGCTGCCCCTCAGTCCGTTGATACCTACTTGCT


                        280                           286
    AsnArgGlnIleAlaGluIleGlyAlaSerLeuIleLysHisTrp
    AATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGA
                   1040                1060                1080                1100
    TTATCTGTCTAGCGACTCTATCCACGGAGTGACTAATTCGTAACCATTGACAGTCTGGTTCAAATGAGTATATATGAAATCT
```

FIG. 3. The sequence of the *amp*^{r} gene. The top strand of sequence is 5′→3′; the bottom strand is 3′→5′, complementary to the top strand. The running count between the strands (every 20 base pairs) starts with the first T in the top strand (the middle of the *Eco*RI site). The three-letter codes for the amino acids of β-lactamase appear directly over their three-base codons and they are numbered (every 10 amino acids) starting from the first methionine.

A few other features of the sequence merit attention, although no supporting biological data currently exist. From positions 189 to 195 is the sequence A-A-T-A-T-T-G, a 5 of 7 match with the canonical "Pribnow box" (21). There is a good correspondence to the complement of the 3′ ends of the 16S ribosomal RNA (22) in positions 199–203. These two regions are likely to represent the recognition sequence for RNA polymerase and the ribosome binding sequence, respectively. Further sequence analysis has located the inverted complementary repeat of the ampicillin transposon Tn3 downstream from the sequence presented here.

The sequence presented in this paper establishes that direct DNA sequence analysis is capable of providing accurate results over extensive regions in a very short time relative to protein sequencing. Fundamentally, DNA sequence analysis can stand on its own merits, but when combined with data about translation or transcriptional products, the results can be essentially infallible.

I thank A. Maxam, P. Farabaugh, A. Hall, D. Hourcade, L. Johnsrud, D. McConnell, W. McClure, R. Ogata, U. Siebenlist, R. Simpson, J. Sims, and R. Tizard for discussions and enzymes. I thank W. Gilbert and J. Knowles for conceiving this exercise and R. P. Ambler and G. K. Scott for their enthusiastic, unselfish participation. This work was supported by National Institutes of Health Grant GM 21514-3 to Walter Gilbert.

1. Knox, J. R., Kelly, J. A., Moews, P. C. & Murthy, N. S., (1976) *J. Mol. Biol.* **104**, 865–875.
2. Hall, A. & Knowles, J. R. (1976) *Nature* **264**, 803–804.
3. Datta, N. and Kontomichalou, P. (1965) *Nature* **208**, 239–241.
4. Meynell, E. & Datta, N. (1967) *Nature* **214**, 885–887.
5. So, M., Gill, R. & Falkow, S. (1975) *Mol. Gen. Genet.* **142**, 239–249.
6. Bolivar, F., Rodriguez, R. L., Betlach, M. C. & Boyer, H. W. (1977) *Gene* **2**, 75–93.
7. Bolivar, F., Rodriquez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L. & Boyer, H. W. (1977) *Gene* **2**, 95–113.
8. Matthew, M. & Hedges, R. W. (1976) *J. Bacteriol.* **125**, 713–718.
9. Tanaka, T. & Weisblum, B. (1975) *J. Bacteriol.* **121**, 354–362.
10. Marinus, M. G. & Morris, N. R. (1975) *Mutat. Res.* **28**, 15–26.
11. Brown, N. L. & Smith, M. (1976) *FEBS Lett.* **65**, 284–287.
12. Maxam, A. M. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
13. Ambler, R. P. & Scott, G. K. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3732–3736.
14. Heffron, F., Sublett, R., Hedges, R. W., Jacob, A. & Falkow, S. (1975) *J. Bacteriol.* **122**, 250–256.
15. Blobel, G. & Dobberstein, B. (1975) *J. Cell. Biol.* **67**, 835–851.
16. Dancer, B. N. & Lampen, J. O. (1975) *Biochem. Biophys. Res. Commun.* **66**, 1357–1364.
17. Farabaugh, P. J. (1978) *Nature*, in press.
18. Roberts, T. M., Shimtake, H., Brady, C. & Rosenberg, M. (1977) *Nature* **270**, 274–275.
19. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M. & Smith, M. (1977) *Nature* **265**, 687–695.
20. Fiers, W. (1975) in *RNA Phages*, ed. Zinder, N. (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York), pp. 353–396.
21. Pribnow, D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 784–788.
22. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.